DOCUMENT RESUME

ED 074 091 TM 002 454

AUTHOR Gullickson, Arlen; Hopkins, Kenneth
TITLE Interval Estimation of Correlation Coefficients from Explicitly Selected Samples.
PUB DATE 71
NOTE 29p.; Ph.D. Thesis, University of Colorado

EDRS PRICE MF-$0.65 HC-$3.29
DESCRIPTORS *Correlation; *Hypothesis Testing; *Mathematical Applications; *Statistical Studies; Technical Reports
IDENTIFIERS Nomograms

ABSTRACT

Two objectives guided the present study: (1) to provide a suitable test for the hypothesis rho=0, and (2) to establish a means by which general users of R can set confidence intervals on rho. The first objective was approached by testing several possible solutions similar to the procedure followed by Forsyth. The second objective was pursued via a combination of a general analytical procedure (Mood and Graybill, 1963) together with computer simulation techniques and a curve fitting technique (Usow, 1970). Procedures for achieving both objectives required the use of R distributions. The method used to obtain the necessary R distributions and the two procedures and their results are described. (Author/DB)

INTERVAL ESTIMATION OF CORRELATION COEFFICIENTS FROM EXPLICITLY SELECTED SAMPLES

Arlen Gullickson, University of Minnesota

and

Kenneth Hopkins, University of Colorado

It is well known that the magnitude of a correlation coefficient is affected substantially by group variability. To adjust for group variability, Pearson (1903) derived a formula for R, a correlation coefficient corrected for restriction of range. The correction preceded the onset of inferential techniques and was developed to estimate ρ for an unrestricted population in situations where complete selection and criterion data were available only from a selected extremity of the population. Essentially, it was a descriptive statistic, provided the assumptions of a bivariate normal population, homoscedasticity of variance errors, and linearity of regression were met. However, R is most often used to estimate ρ when both selection and criterion data are available only from a sample of a selected extremity of the population. Consequently, although R was developed for descriptive purposes, its current use is primarily as an inferential statistic.

As a descriptive statistic, R will not equal ρ only if the underlying assumptions have not been met. Two empirical studies (Hovis, 1935; Creager, 1953) illustrated that in several cases involving large N's the assumptions are met well enough so that R describes ρ very well. The same situation does not hold for inferential statistics. Due to the presence of sampling

error, one does not expect a statistic to equal the parameter even in cases where the underlying assumptions have been exactly met. For this reason, R is a meaningful inferential statistic only to the extent that an investigator can specify boundaries within which ρ will lie. Presently, investigators cannot determine the accuracy of corrected correlation coefficient, ρ, because its inferential characteristics have not been assessed. As a result, Lord and Novick (1968, p. 147) stated, ". . . a more cautious attitude toward these formulas is called for in any applications in which the ratio of standard deviations in the unselected group to standard deviations in the selected group is more than 1.40. This condition corresponds to a selection of approximately the upper 70 percent from a standard normal population."

It has been difficult to explicate the properties of R primarily because R is dependent upon three parameters; sample size (N), the correlation coefficient between the two variables X and Y in the unrestricted population (ρ), and the percentile point of the X variable such that all X values included in the explicitly selected sample are larger than A $[P(A)]$. The interdependency among these factors causes intractable mathematical problems that have thus far precluded an analytical solution to the density function of R.

Only one study (Forsyth, 1971) has been undertaken to clarify the inferential properties of R. Using computer simulation methods to simply test the efficacy of a hypothesized solution, Fisher's log transformation of R to set confidence intervals on ρ, it was established that the procedure does not produce suitable accurate confidence intervals. An

attempt to "correct" the formula by adjusting the degrees of freedom in the Z statistic improved the results, but did not provide a definitive solution to the interval estimation problem.

Two objectives guided the present study: (1) To provide a suitable test for the hypothesis $\rho=0$, and (2) To establish a means by which general users of R can set confidence intervals on ρ. The first objective was approached by testing several possible solutions similar to the procedure followed by Forsyth. The second objective was pursued via a combination of a general analytical procedure (Mood & Graybill, 1963) together with computer simulation techniques and a curve fitting technique (Usow, 1970). Procedures for achieving both objectives required the use of R distributions. The method used to obtain the necessary R distributions is described immediately below. Following that, the two procedures and their results are described.

There are two alternative formulas that can be used to obtain R values, and hence estimate ρ. One is in common use, a description of it can be found in Gulliksen (1950) and several other sources. A second formula described by Kelley (1923) yields approximately the same value of R as does the conventional formula but is more difficult to use and generally resulted in less acceptable tests of $\rho=0$ (Gullickson, 1971). For those reasons, only the procedures and results as they pertain to the conventional formula are described in this paper. (Lower case letters denote values from the restricted group; capital letters indicate values for the unrestricted group.)

$$R = \frac{S_X r}{\sqrt{s_x^2 - s_x^2 r^2 + S_X^2 r^2}}$$

Forsyth (1971) found R distributions obtained by using σ_X in the above formula differed little from R distributions obtained by using S_X, hence, σ_X was used in the place of S_X for the computation of all R values since it simplified the procedure and reduced computer costs.

Values of r and s_x were obtained using a set of normal deviates, N(0,1) (Collins, 1970) toge... with a random number generator (Jordan, 1970) and a computer simulation method for obtaining correlated pairs from a population of paired variates having a correlation of ρ (Lehman & Bailey, 1968, p. 228). In this process, X variates of the (X,Y) pairs were always randomly selected from the population of X values greater than P(A). Each r value thus obtained was then corrected for explicit selection via the conventional formula to obtain R. The procedure was replicated, holding N, ρ, and P(A) constant, a designated number of times to produce a distribution of R sample point estimates of ρ.

Testing the Hypothesis $\rho=0$

As indicated in Table 1, R distributions, each composed of 1,000 sample points, were formed for $\rho=0$, N=27, 52, 100, and for P(A)=.10, .50, .75.

Insert Table 1 about here

Five test statistics were applied to the R sample points in each of the nine distributions in an attempt to find the adequacy of the test statistics in terms of actual significance levels being equal to respective nominal significance levels. The five test statistics are listed below:

TABLE 1

Number of Replications Used for Building R

Distributions of Estimates of ρ When $\rho=0$

N	Number of R Sample Points Per Distribution	P(A)
27	1,000	.10, .50, .75
52	1,000	.10, .50, .75
100	1,000	.10, .50, .75

(1) $$z = \frac{R}{\sigma_R} \quad \text{where} \quad \sigma_R = 1/\sqrt{N-1}$$

(2) $$z = \frac{R}{\sigma_R} \quad \text{where} \quad \sigma_R = \frac{1}{\sqrt{N-1}}\left(\frac{R(1-R^2)}{r(1-r^2)}\right)$$ Kelley (1923, p. 316)

(3) $$t = \frac{R}{S_R} \quad \text{where} \quad S_R = \sqrt{\frac{1-R^2}{N-2}}$$

(4) $$t = \frac{R}{S_R} \quad \text{where} \quad S_R = \sqrt{\frac{1-r^2}{N-2}}\left(\frac{R(1-R^2)}{r(1-r^2)}\right)$$ Kelley (1923, p. 316)

(5) $$z = \frac{Z}{\sigma_Z}$$ where Z is the Fisher log transformation of R and $\sigma_Z = 1/\sqrt{N-3}$

Of the five formulas 1, 3, and 5 were completely unacceptable: the actual Type I error probabilities substantially exceeded the nominal significance levels. Formulas 2 and 4, however, proved to be more accurate. Although formula 2 yields a z statistic, and formula 4 yields a t statistic, both formulas utilize the Kelley formula to obtain the standard error of R, and the results obtained from the two procedures were quite comparable. Both exhibited a similar trend and in no case did one appear to be significantly better than the other. The average difference in Type I error probability between the results of the two formulas was only .0014. Because the two formulas are so similar only the results for formula 4 are included here (see Table 2). Results for formulas 1, 2, 3, and 5 are given by Gullickson (1971).

Insert Table 2 about here

TABLE 2

Actual Probability of a Type I Error for Testing

the Hypothesis $\rho=0$ with a Two-Tailed t-Test

for Various Sample Sizes and Explicit Selection Points

$$t = \frac{\sqrt{r \quad (1-r^2) \quad (N-2)}}{1-R^2}$$

Nominal Probability of Type-I Error		.01	.05	.10	.20
P(A)[a]	N	Actual Probabilities of Type-I Error			
.10	27	.026	.063	.115	.207
.10	52	.014	.054	.099	.171
.10	100	.008	.052	.098	.206
.50	27	.048	.099	.155	.219
.50	52	.021	.069	.112	.181
.50	100	.016	.059	.107	.207
.75	27	.069	.135	.180	.246
.75	52	.032	.086	.130	.198
.75	100	.022	.069	.119	.223

[a] 1-P(A) is the proportion of the unrestricted sample employed.

In general, formula 4 provides a liberal test of α, i.e., the actual chance of a Type I error is greater than the stated nominal level. The discrepancy between the actual and nominal significance levels becomes less pronounced as any one or combination of the following occur: (1) α is increased in magnitude, (2) N is increased in size, (3) P(A) is reduced. For example, when N=27 and α=.01 the estimated actual significance level decreased from .069 to .026 as P(A) was reduced from .75 to .10. For the same α but N=100, the estimated actual significance level was .022 when P(A)=.75 but reduced to .008 when P(A)=.10. For research purposes, it is recommended that either formula 2 or 4 be employed for hypothesis testings purposes, but that no test be made if both N and the proportion in the explicitly selected sample are small.

Interval Estimation on ρ

Since analytic means of setting confidence intervals are not available, four sample sizes N=25, 50, 100, and 200, six explicit selection points such that P(A)=.10, .20, .40, .60, .75, and .90, and ten correlations ρ=0, .1, .2, · · ·, .9 were used in all possible combinations to produce a total of 240 R distributions (see Table 3).

Insert Table 3 about here

Those 240 distributions were in turn used to build 24 confidence interval nomograms for each of two α values (α=.01 and .05). All nomograms for a set combination of α and P(A) were then placed on one R vs. ρ axis to provide a total of 12 sets of nomograms as provided in Figures 2-13.

To build a single 1-α confidence interval nomogram, the ten R distributions for ρ=0, .1, .2, · · ·, .9 and a single combination of N and P(A)

TABLE 3

Number of Replications Used in Building Each

R Distribution for Interval Estimation on ρ

Sample Size N	P(A)	ρ	Number of Replications per Distribution
25	.10, .20, .40, .60, .75, .90	0, .1, .2, · · ·, .9	10,000
50	.10, .20, .40, .60, .75, .90	0, .1, .2, · · ·, .9	6,000
100	.10, .20, .40, .60	0, .1, .2, · · ·, .9	1,500
100	.75, .90	0, .1, .2, · · ·, .9	3,000
200	.10, .20, .40, .60	0, .1, .2, · · ·, .9	750
200	.75, .90	0, .1, .2, · · ·, .9	1,500

Note: Each distribution was formed using a single combination of N, P(A) and ρ e.g., using the combination N=25, P(A)=.10, and ρ=0, 10,000 replications were made to form an R distribution.

were used (e.g., the ten ρ values for N=25 and P(A)=.10). Nine of the ten R distributions, those derived under the conditions ρ=.1, .2, .3, · · ·, .9, were essentially used twice. Because ρ is symmetrical about zero, each R value of the distributions could be multipled by -1 to produce new distributions corresponding to ρ=-.1, -.2, -.3, · · ·, -.9.

The lower bound of each confidence interval nomogram was formed by determining the $\alpha/2$ percentile point of each of the 19 R distributions, pairing each with the ρ value it estimated, and using those 19 number pairs with the two additional (R,ρ) pairs (-1,-1) and (1,1) to derive a polynomial line of best fit. The line forming an upper confidence interval bound was obtained in the same manner except the $1-\alpha/2$ percentile points of the R distributions were used. Those two lines on an R vs. ρ axis form confidence interval bounds on ρ. Figure 1, an illustrative confidence interval nomogram, allows one to set a .95 confidence interval on ρ for an R obtained from an explicitly selected sample of N=25 and P(A)=.1. If under the specified conditions, a user obtained an R value of .5 the confidence interval on ρ would have lower and upper bounds of -.02 and .75 respectively.

Insert Figure 1 about here

A single $1-\alpha$ nomogram provides precise confidence interval on ρ only for a set combination of N and P(A). Obviously there are an infinite number of such combinations, and no single nomogram would exactly fit more than a few cases. However, the combination of four nomographs on a single axis allows a user to interpolate and set confidence intervals

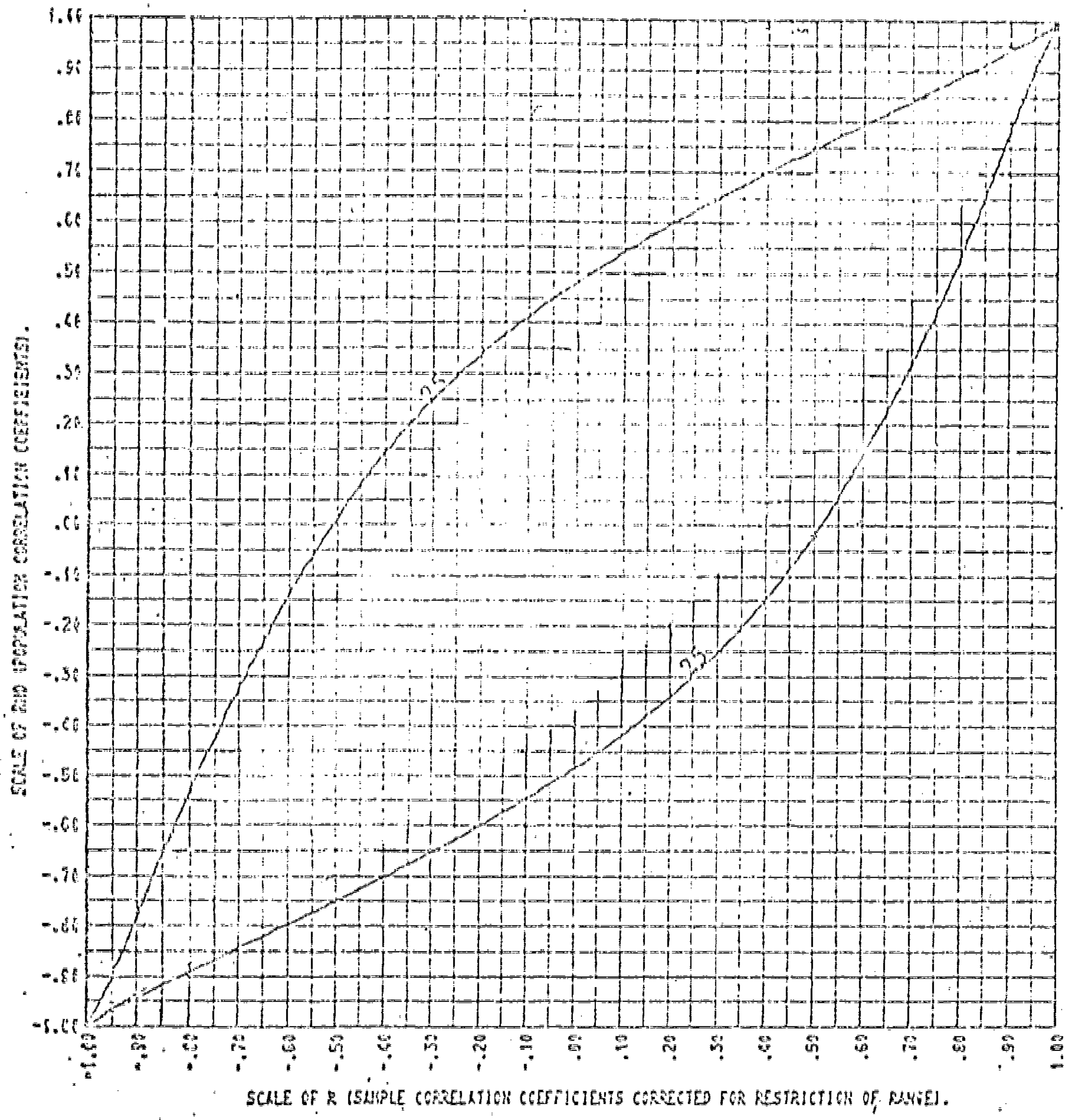

Fig. 1. The 95% confidence intervals around R, corrected for restriction of range, on ρ for N = 25, when P(A) = .10. (Find the upper limit value above the Principal diagonal and the lower limit value below it.)

for a wide range of sample sizes. Also, the series of Figures 2-7, and 8-13, allow a user to interpolate across values of P(A). By interpolating within and across figures, a user can set confidence intervals on ρ for any R regardless of the sample size or the cutoff point used for explicit selection.

As can be noted from Figures 2-13, the variance of R increases as P(A) is increased. That phenomenon appears to be the cause of an increasing amount of error in the polynomial lines of best fit (Figures 2-13) as P(A) gets large. Because the variance of the R distribution is much larger for P(A)=.90 than for P(A)=.10, the precision with which the $\alpha/2$ and $1 - \alpha/2$ points of the R distribution were located was correspondingly decreased. As is noted in Table 3, a very large number of sample points per distribution was obtained for all sample sizes when P(A)=.75 and .90 in an attempt to overcome that problem. The following empirical check illustrates that the errors, though large for P(A)=.90, do not materially reduce the precision of the respective confidence interval nomograms.

As a check on the empirically obtained confidence intervals a simulation of 10,000 replications was run for N=25, P(A)=.90, and ρ=.65, .75, and .85. The obtained points were placed in their respective positions for the lower line on Figure 7. The largest difference between an obtained point and the line, .08, occurred when ρ=-.75 on Figure 7. Note that if that empirical checkpoint were used instead of the line, it would result in a confidence interval, on ρ, being longer by only approximately .03 on the lower tail and .02 on the upper tail. That corresponds to an error of 3.3 percent in the confidence interval (total error, .05, divided by total confidence interval width, 1.5). Obviously as the polynomial lines approach vertical,

any error would produce a large percentage error in terms of the total confidence interval length. However, as can be seen, the difference between the checkpoints and the line is negligible in the region of the large slope.

Insert Figures 2-13 about here

As was noted at the beginning of this article, an indicator of "goodness" for an inferential statistic is the width of its resulting confidence interval on the desired parameter. The width of R's confidence interval on ρ are very dependent on both P(A) and N. The relationship, although visible in Figures 2-13, may be seen more clearly in Figure 14. To obtain Figure 14, the 95 percent confidence intervals were measured for set R and N values (Figures 8-13) and then plotted against the respective P(A) values. (Confidence interval widths at P(A)=0 were obtained from a table of confidence intervals about r on ρ (Glass & Stanley, 1970, p. 537) because when P(A)=0, R is a true Pearson r.)

Insert Figure 14 about here

The relationships illustrated by Figure 14 can be summarized in four generalizations:

1. For constant P(A) and R, as N is increased in size the confidence interval decreases.
2. As P(A) is increased, i.e., the proportion in the explicitly selected sample decreases, the confidence interval width increases proportionately. In every case the series of points for a set combination of R and N indicated a linear relationship between confidence interval width and P(A).

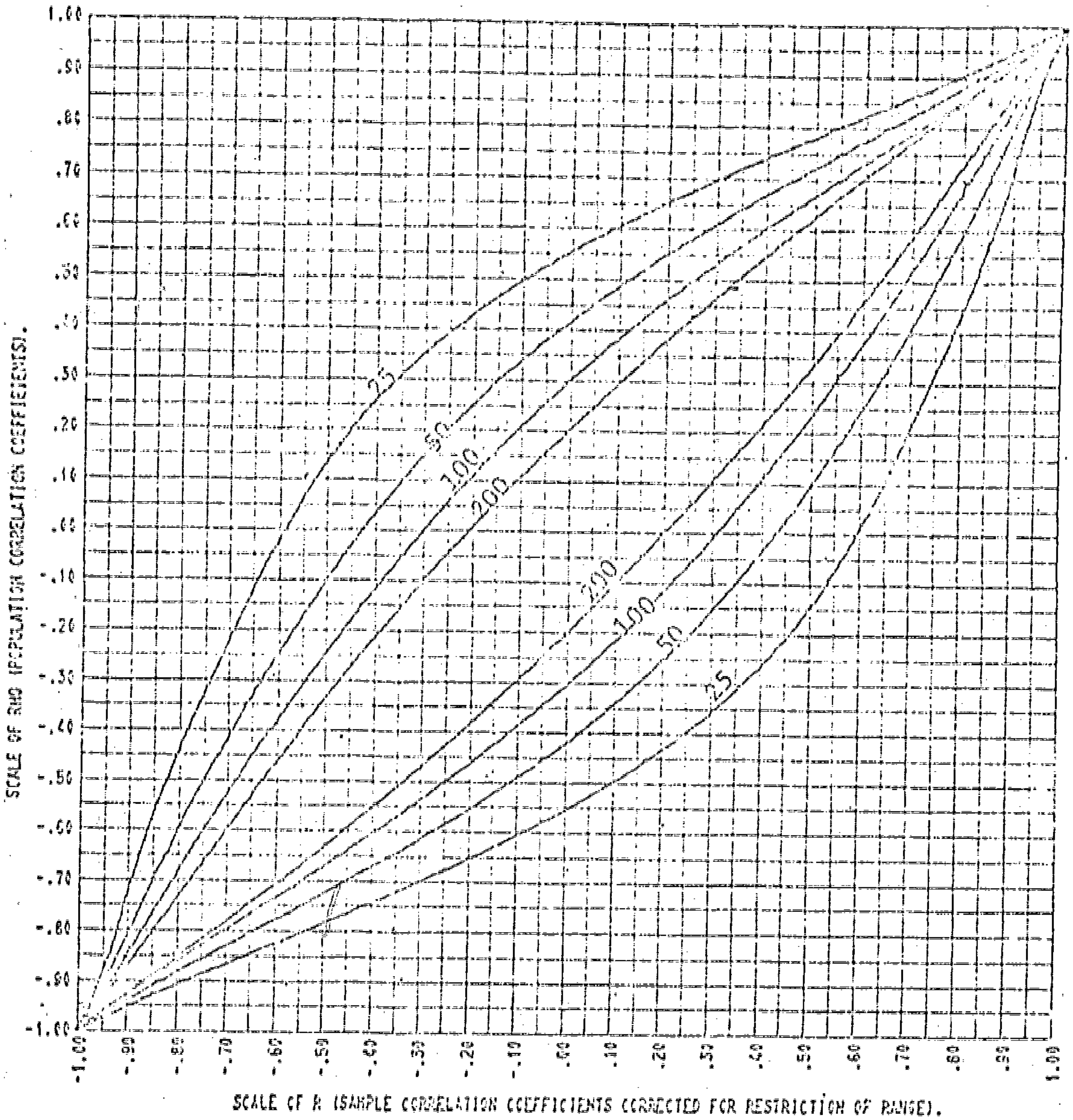

Fig. 2. The 99% confidence intervals around R, corrected for restriction of range, on ρ for N = 25, 50, 100, and 200 when P(A) = .10. (Find the upper limit value above the Principal diagonal and the lower limit value below it.)

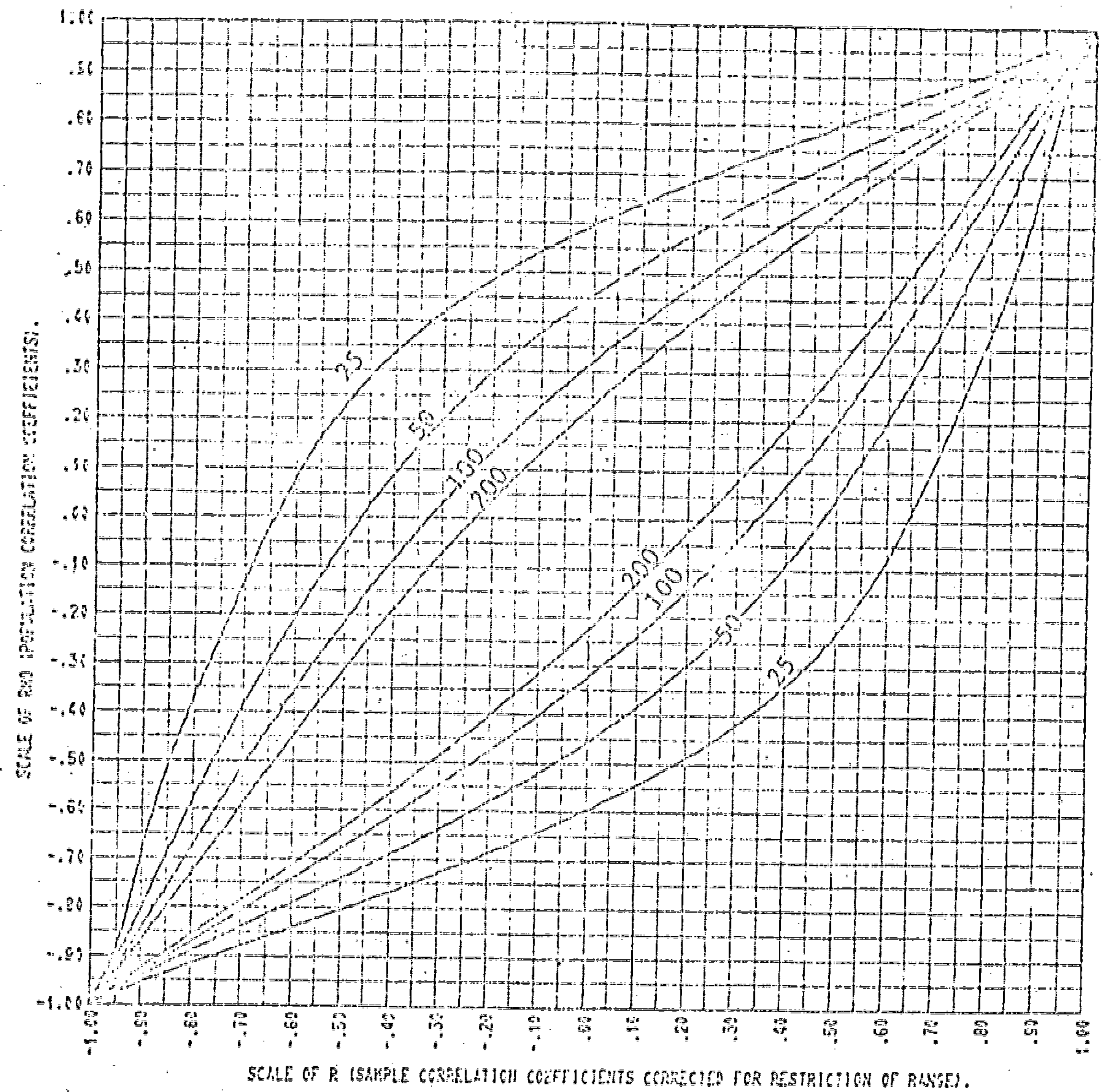

Fig. 3. The 99% confidence intervals around R, corrected for restriction of range, on ρ for N = 25, 50, 100, and 200 when P(A) = .20. (Find the upper limit value above the Principal diagonal and the lower limit value below it.)

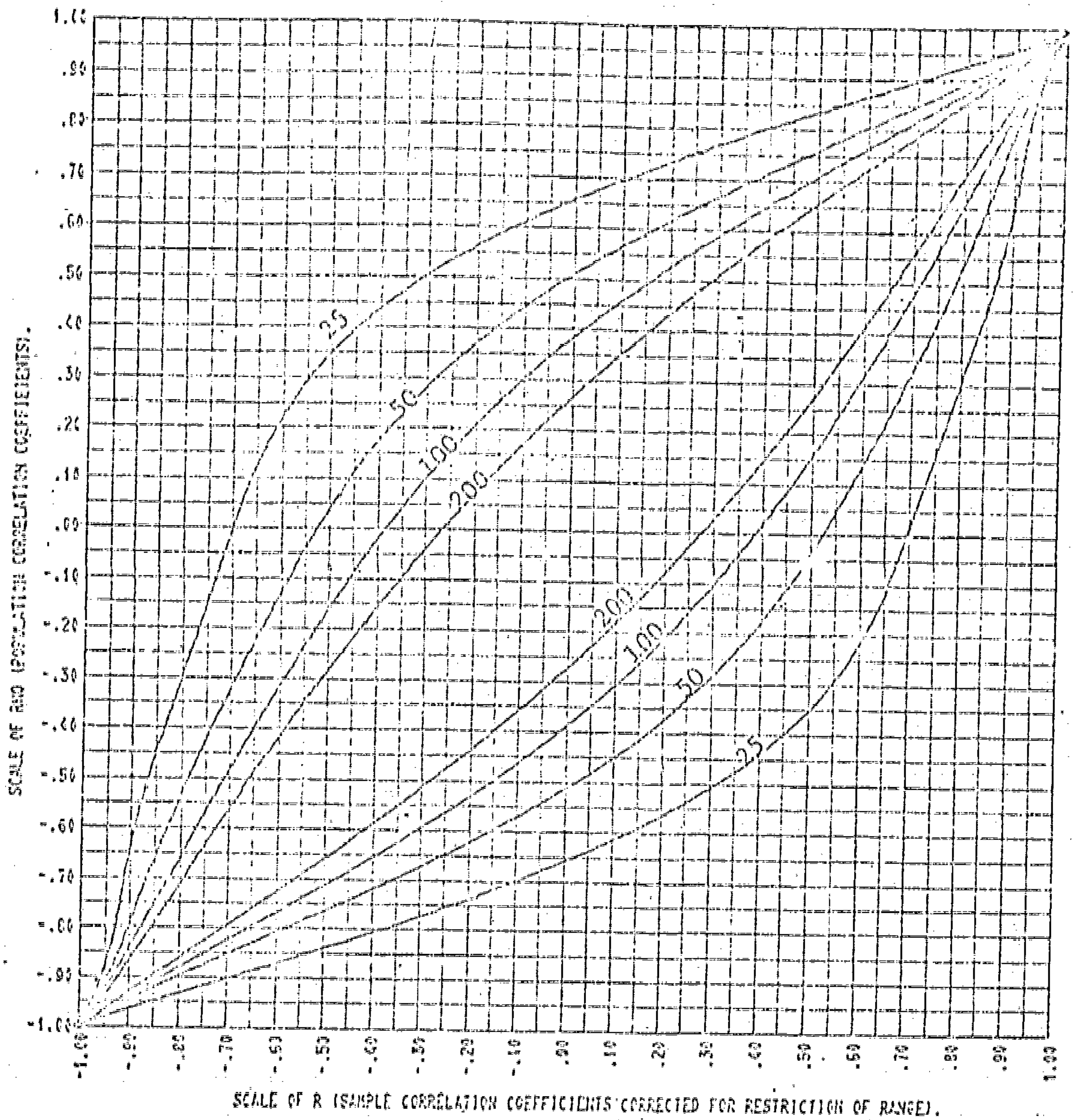

Fig. 4. The 99% confidence intervals around R, corrected for restriction of range, on ρ for N = 25, 50, 100, and 200 when P(A) = .40. (Find the upper limit value above the Principal diagonal and the lower limit value below it.)

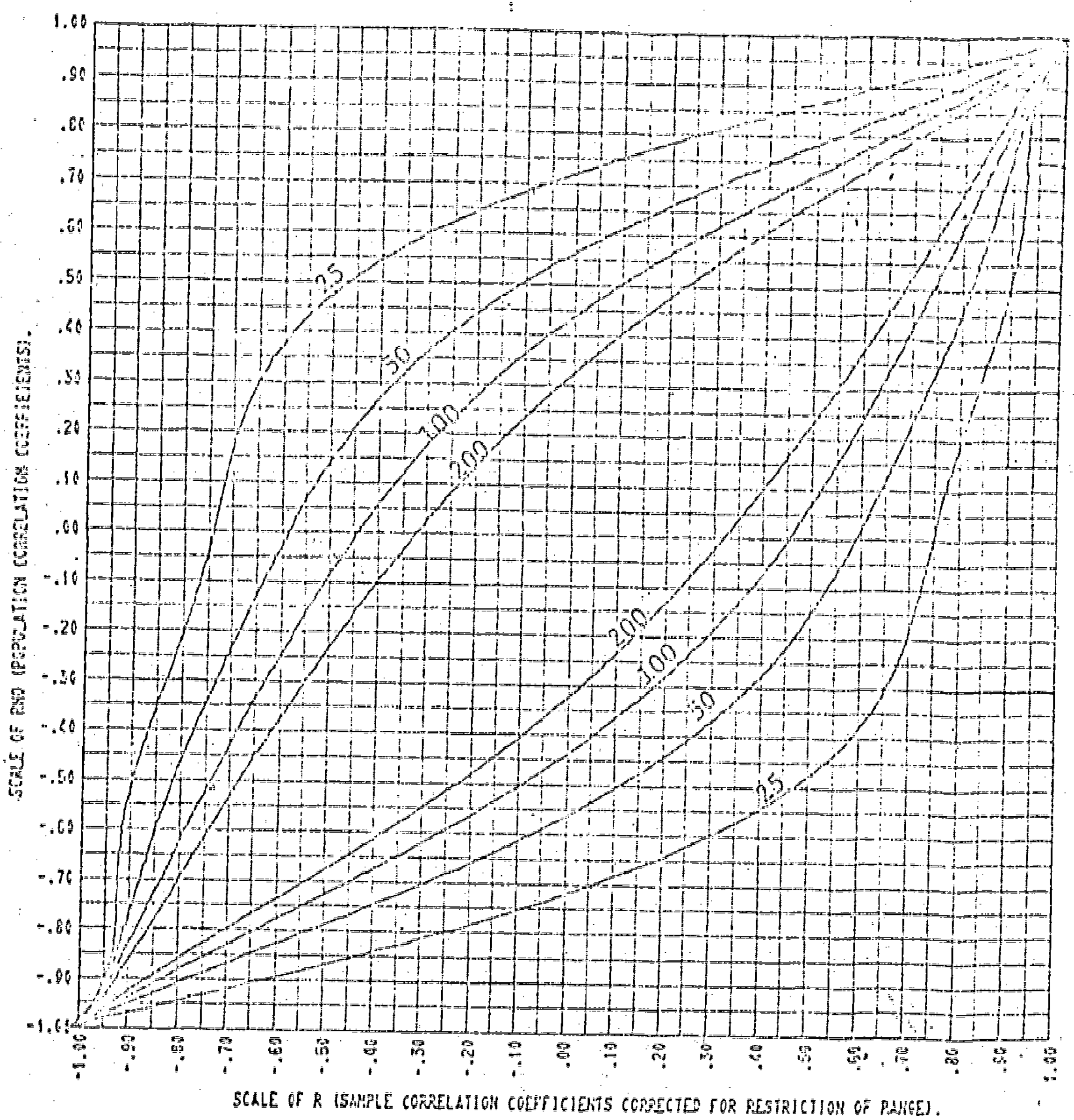

Fig. 5. The 99% confidence intervals around R, corrected for restriction of range, on ρ for N = 25, 50, 100, and 200 when P(A) = .60. (Find the upper limit value above the Principal diagonal and the lower limit value below it.)

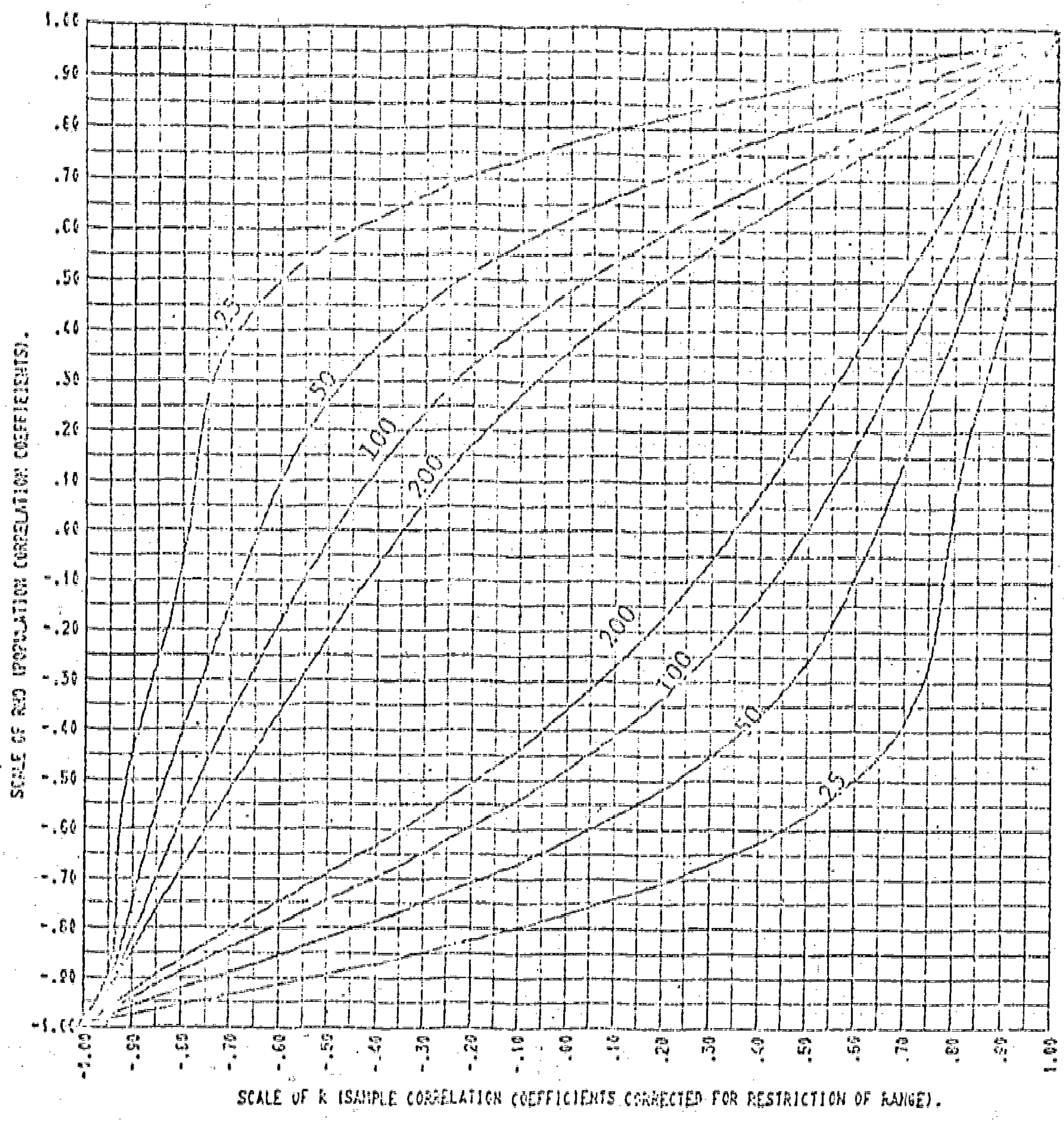

Fig. 6. The 99% confidence intervals around R, corrected for restriction of range, on ρ for N = 25, 50, 100, and 200 when P(A) = .75. (Find the upper limit value above the Principal diagonal and the lower limit value below it.)

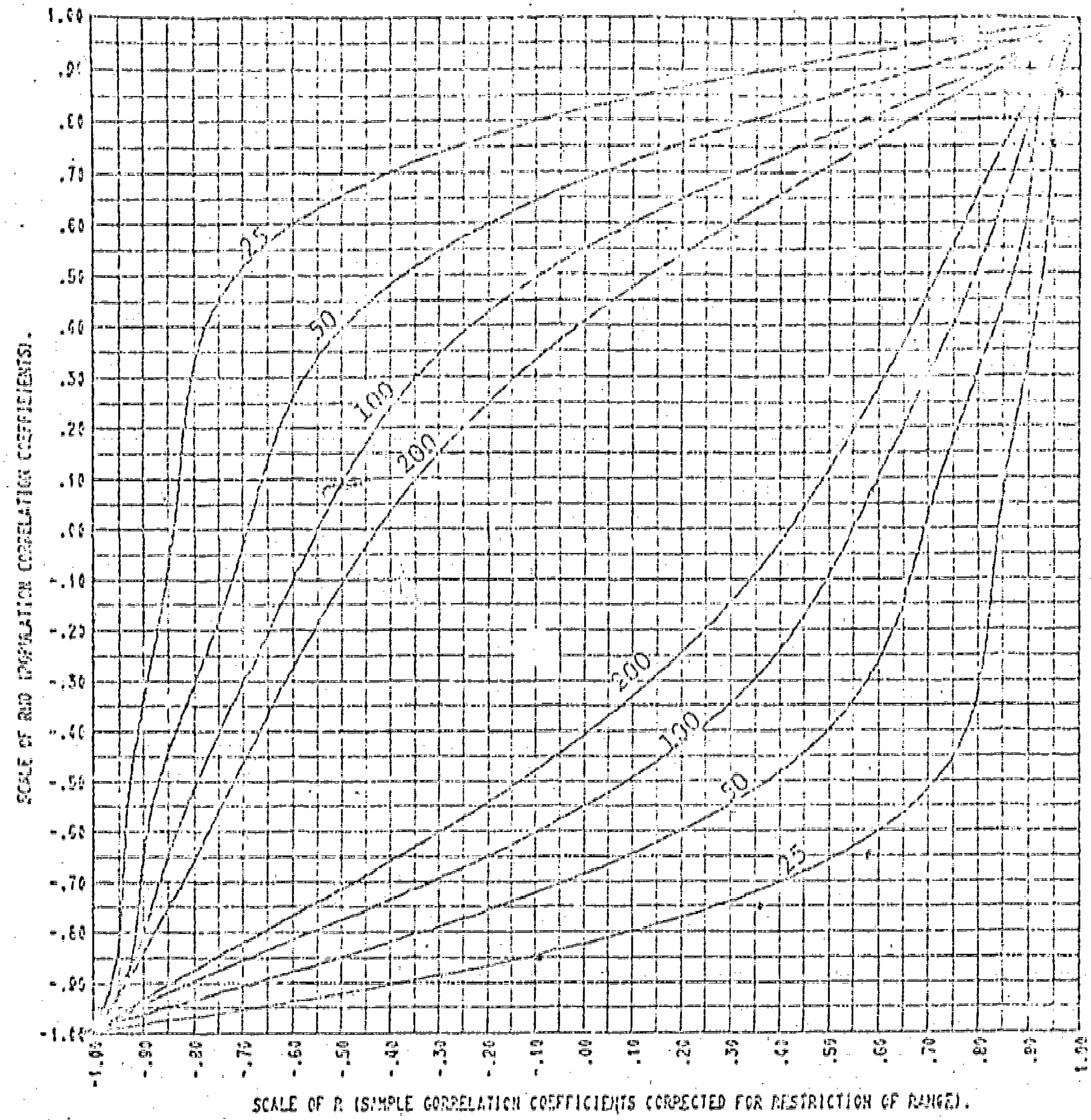

Fig. 7. The 99% confidence intervals around R, corrected for restriction of range, on ρ for N = 25, 50, 100, and 200 when P(A) = .90. (Find the upper limit value above the Principal diagonal and the lower limit value below it.)

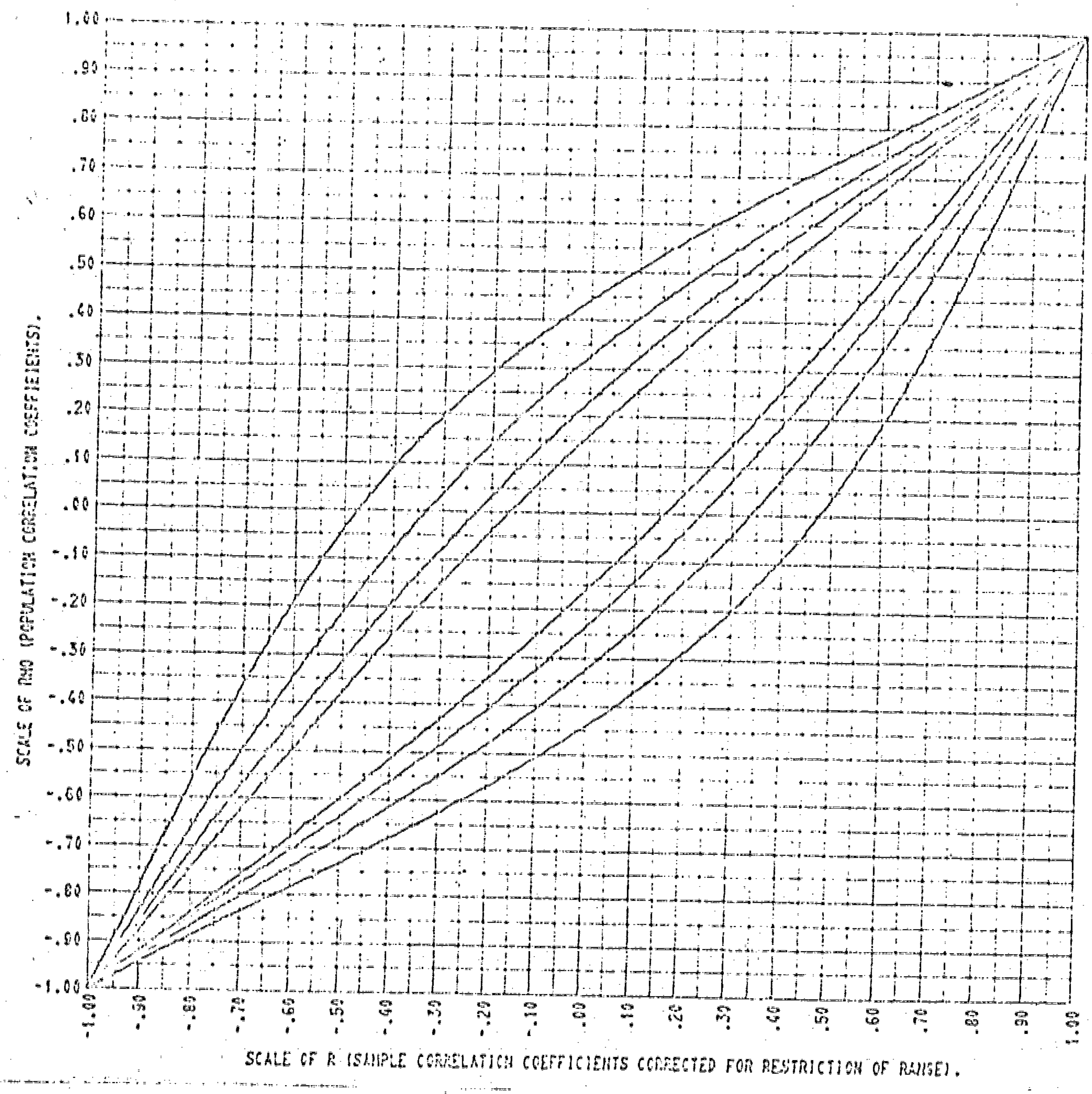

Fig. 8. The 95% confidence intervals around R, corrected for restriction of range, on ρ for N = 25, 50, 100, and 200 when P(A) = .10. (Find the upper limit value above the Principal diagonal and the lower limit value below it.)

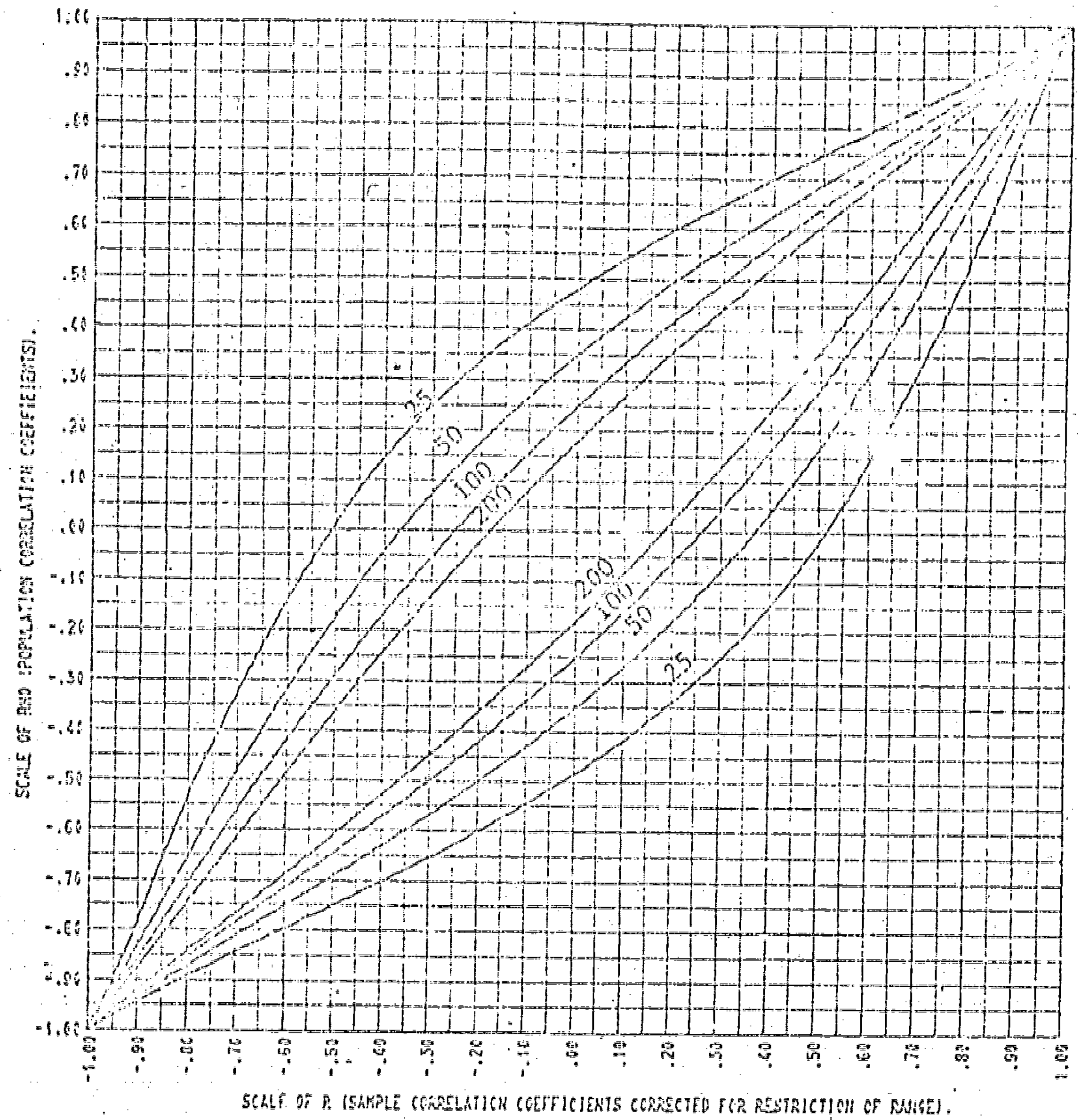

Fig. 9. The 95% confidence intervals around R, corrected for restriction of range, on ρ for N = 25, 50, 100, and 200 when P(A) = .20. (Find the upper limit value above the Principal diagonal and the lower limit value below it.)

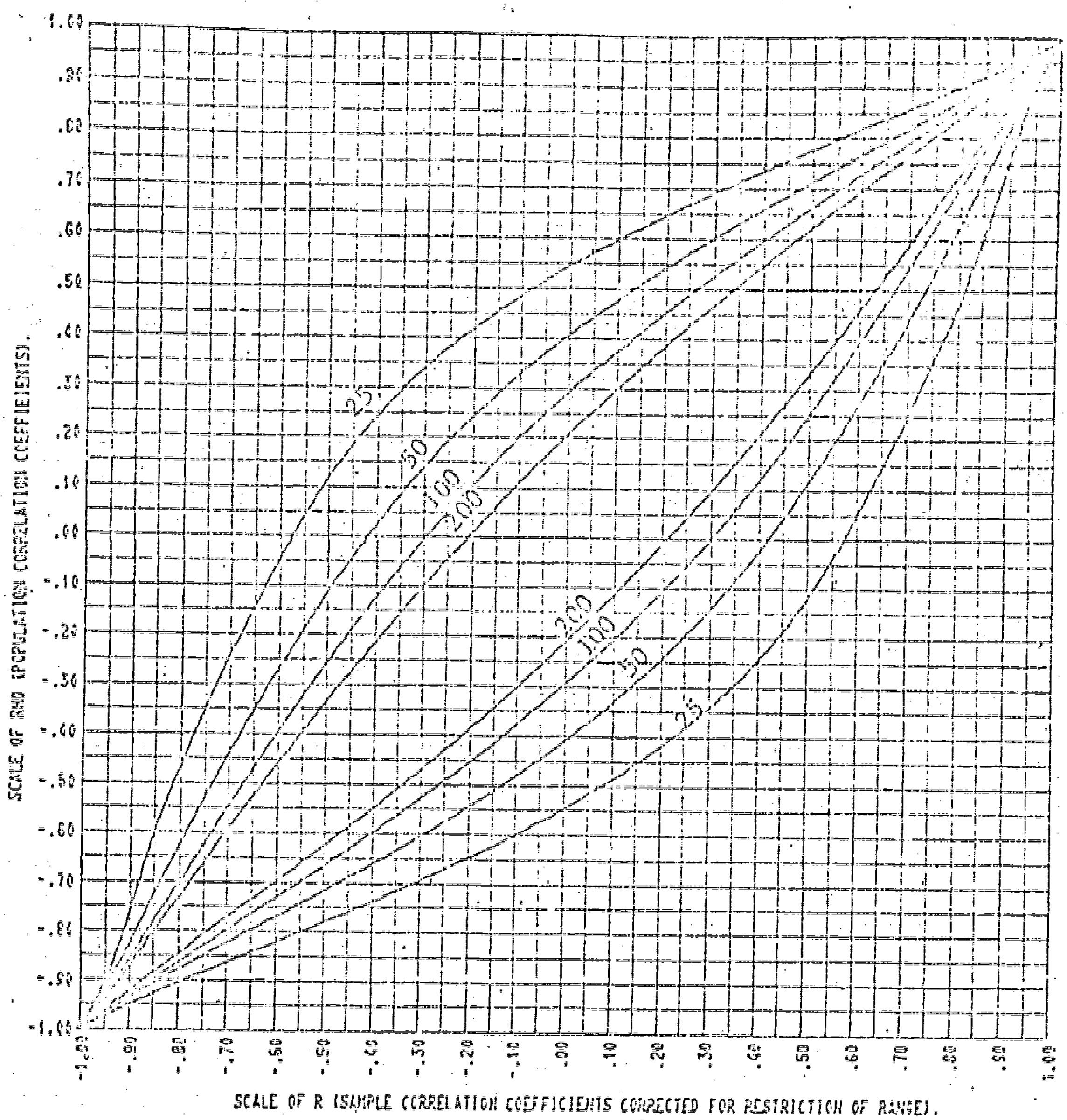

Fig. 10. The 95% confidence intervals around R, corrected for restriction of range, on ρ for N = 25, 50, 100, and 200 when P(A) = .40. (Find the upper limit value above the Principal diagonal and the lower limit value below it.)

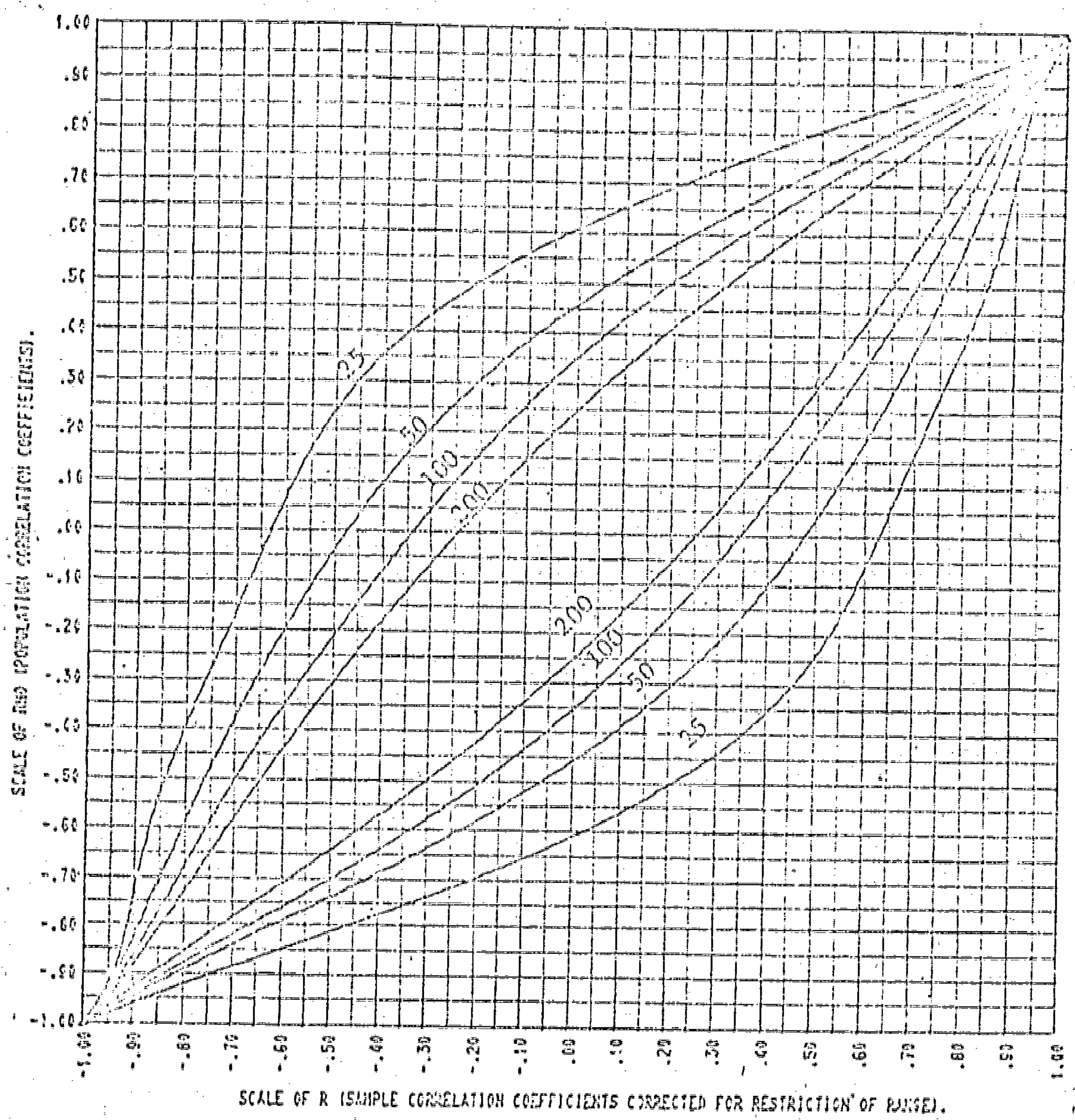

Fig. 11. The 95% confidence intervals around R, corrected for restriction of range, on ρ for N = 25, 50, 100, and 200 when P(A) = .60. (Find the upper limit value above the Principal diagonal and the lower limit value below it.)

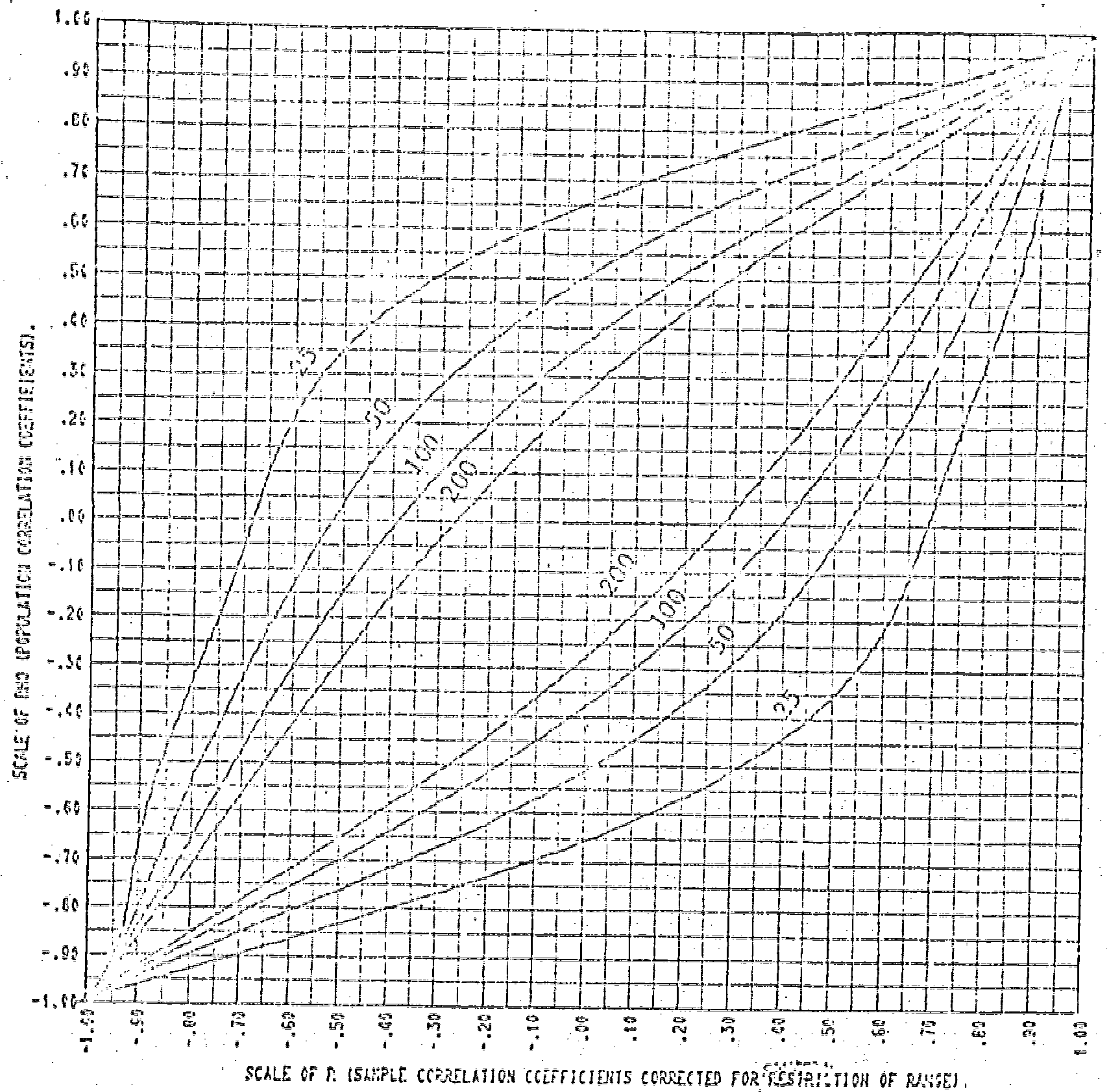

Fig. 12. The 95% confidence intervals around R, corrected for restriction of range, on ρ for N = 25, 50, 100, and 200 when P(A) = .75. (Find the upper limit value above the Principal diagonal and the lower limit value below it.)

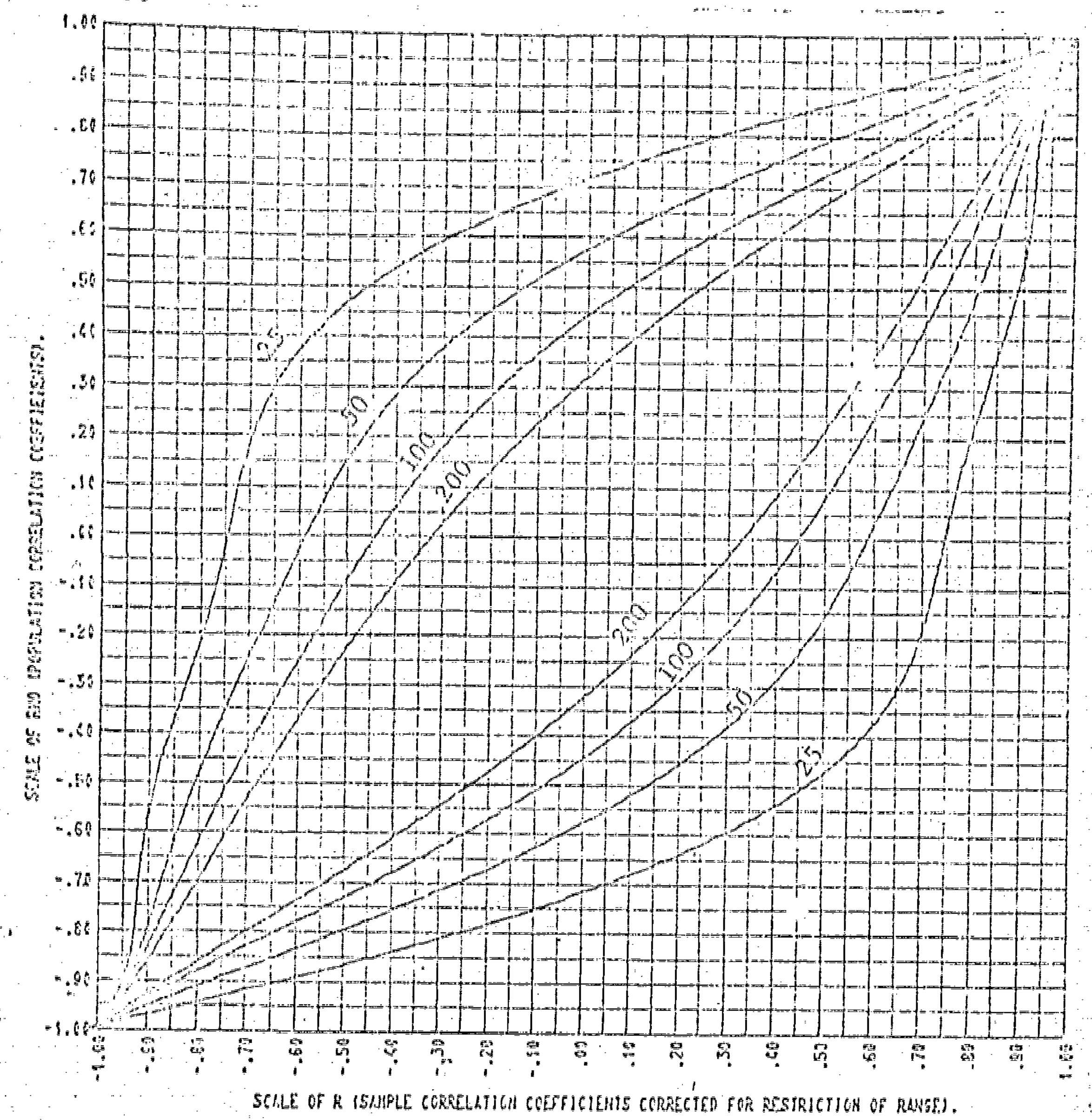

Fig. 13. The 95% confidence intervals around R, corrected for restriction of range, on ρ for N = 25, 50, 100, and 200 when P(A) = .90. (Find the upper limit value above the Principal diagonal and the lower limit value below it.)

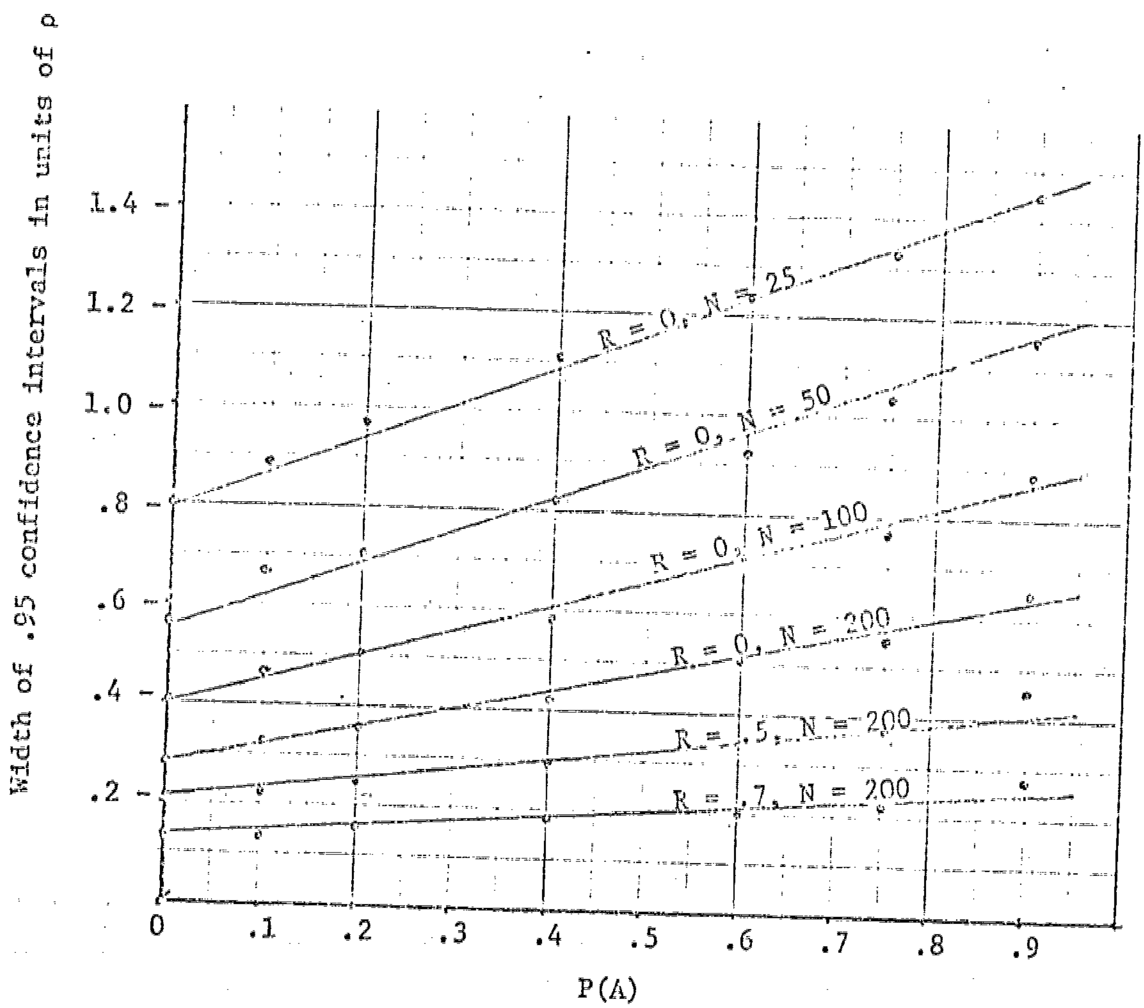

Fig. 14. A graphic representation of the effects of P(A), N, and R on confidence interval width for .95 confidence intervals on ρ.

3. The rate of change in confidence interval width per unit change in P(A) (the slope of the line) is dependent upon N and decreases as N is increased in size. Note the top four lines of Figure 14. R=0 for each of the four lines, and as N increases from N=25 for the top line to N=200 for the fourth line, the slope gradually but noticeably decreases.

4. The rate of change in confidence interval width per unit change in P(A) is dependent on R and decreases as R increases in size. Note the bottom three lines of Figure 14. For each of the three lines with N=200, with R=0, R=.5, and R=.7, note that as R is increased, the slope of the line decreases.

Points 2, 3, and 4 make it clear that when there is explicit selection on one variable, a considerable price is exacted in terms of the precision with which inferential statements about ρ can be made. The greater P(A) becomes, the greater will be the corresponding loss of precision. By increasing N the loss of precision caused by increasing P(A) can be reduced; also, the loss of precision per unit change in P(A) is decreased as R increases. However, it appears that only as N becomes very large and R approaches ± 1 will the effects of selection be negligible.

One additional point should be noted. Because of the large confidence intervals on ρ, when N is small and P(A) is large, the obtained R may have little practical value, other than to prevent

over-interpretation of sample R values. For example, when N=25, and P(A)=.90, not until |R| is greater than .84 can ρ be said to be different from zero at the 99 percent level of confidence.

Certainly, characteristics just described represent a very real improvement in our knowledge of the inferential properties of R. For the user, who desires to estimate ρ from an explicitly selected sample, the nomograms (Figures 2-13) provide an efficient and accurate means of setting those confidence intervals once R has been calculated. It is obvious the nomograms are not the ultimate elegant solution, a simple confidence interval formula would be much better; but the ease with which they can be applied makes them a viable aid to most practitioners.

BIBLIOGRAPHY

Collins, J. R., Jackknifing Generalizability. (unpublished Ph.D. Thesis) University of Colorado, 1970.

Creager, J., Studies in Methodology II. Efficacy of the Univariate Formulas for Correcting for Restriction of Range. Human Resources Res. Cent. Staff Res. Memorandum, mimeographed, 1953. Cited by S. Rydberg, Bias in Prediction. Stockholm: Almquist & Wiksell, 1963.

Forsyth, R. A., An Empirical Note on Correlation Coefficients Corrected for Restriction in Range. Educational and Psychological Measurement, 1971, 31, pp. 115-123.

Glass, G. V. & Stanley, J. C., Statistical Methods in Education and Psychology. New Jersey: Prentice Hall, Inc., 1970.

Gullickson, A. R., Interval Estimation of Correlation Coefficients from Explicitly Selected Samples. (unpublished Ph.D. Thesis) University of Colorado, 1971.

Gulliksen, H., Theory of Mental Tests. New York: John Wiley and Sons, 1950.

Hovis, R. S., An Evaluation and Comparison of Two Formulas for Correcting Coefficients of Correlation for Heterogeneity. Master's Thesis, Pennsylvania State College, 1935.

Jordan, H. E., URNZ, Computer Program, Graduate School Computing Center, University of Colorado, Boulder, Colorado, 1970.

Kelley, T. L., Statistical Methods. New York: Macmillan, 1923.

Lehman, R. S., & Bailey, D. E., Digital Computing: Fortran IV and Its Applications in Behavioral Science. New York: John Wiley and Sons, 1968.

Lord, R. M., & Novick, M. R., Statistical Theories of Mental Test Scores. Reading, Mass.: Addison-Wesley, 1968.

Mood, A. M., & Graybill, F. A., Introduction to the Theory of Statistics. (2nd ed.). New York: McGraw-Hill, Inc., 1963.

Pearson, K., Mathematical Contributions to the Theory of Evaluation-XI. On the Influence of Natural Selection on the Variability and Correlation of Organs. Royal Society of London, Philosophical Transactions, London Series A, 1903, 200, pp. 1-66.

Usow, K. PLSQZ, Computer Program, Graduate School Computing Center, University of Colorado, Boulder, Colorado.